面向人脸视频防伪检测的大规模中文数据评测基准*

贝毅君[1,2]，娄恒瑞[1]，高克威[1]，宋杰[1,2]，王蕊[3]，金苍宏[4]，宋明黎[2]，冯尊磊[1,2]

[1](浙江大学 软件学院,浙江 宁波 315103)

[2](浙江大学 计算机科学与技术学院,浙江 杭州 310007)

[3](中国科学院 信息工程研究所,北京 100093)

[4](浙大城市学院 计算机与计算科学学院,浙江 杭州 310015)

通讯作者: 冯尊磊, E-mail: zunleifeng@zju.edu.cn

摘 要:随着 AIGC 技术的快速发展,逼真的伪造人脸视频已经可以欺骗人类视觉感知.因此,大量人脸防伪检测算法被提出用于伪造人脸视频的检测.然而如何有效评估这些伪造检测算法的有效性与可应用性,仍面临着诸多挑战.为有效推动人脸防伪检测成效的量化评估与防伪检测技术迭代发展,本文提出了一项面向人脸视频防伪检测的大规模中文数据评测基准,发布了全球首个 CHN-DF 中文数据集(https://github.com/HengruiLou/CHN-DF).填补了人脸视频防伪数据集在大规模中文数据方面的空白.本文详细介绍了构建 CHN-DF 数据集和中文数据评测基准的流程,并通过实验验证了 CHN-DF 数据集的复杂性和贴近真实场景水平.期望该评测基准能帮助研究人员构建更实用有效的人脸视频防伪检测模型,推动防伪检测领域技术发展.同时,本文指出了中文人脸视频防伪检测基准数据集和防伪检测技术所面临的挑战,提出了未来可能的研究方向,为推动人脸视频防伪检测技术发展提供了有益思路.

关键词: 深度学习;深度伪造;假视频;多模态防伪检测

中图法分类号: TP393

Large-Scale Chinese Data Benchmark for Face Video Anti-Forgery Identification

BEI Yi-Jun[1,2], LOU Heng-Rui[1], GAO Ke-Wei[1], SONG Jie[1,2], WANG Rui[3], JIN Cang-Hong[4], SONG Ming-Li[2], FENG Zun-Lei[1,2]

[1](College of Software Technology, Zhejiang University, Ningbo 315003, China)

[2](College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China)

[3](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

[4](College of Computer and Computer Science, Hangzhou City University, Hangzhou 310015, China)

Abstract: With the rapid development of AIGC (Artificial Intelligence Generated Content) technology, hyper-realistic forged facial videos have become capable of deceiving human visual perception. As a result, a significant number of facial anti-forgery detection algorithms have been proposed for the identification of these fake facial videos. However, effectively evaluating the efficacy and applicability of these forgery detection algorithms remains a substantial challenge. To effectively promote the quantitative assessment of facial anti-forgery detection performance and the iterative development of anti-forgery technologies, this paper introduces a large-scale Chinese data benchmark for facial video anti-forgery identification and releases the world's first CHN-DF Chinese dataset (https://github.com/HengruiLou/CHN-DF), filling the gap in facial video anti-forgery datasets in terms of large-scale Chinese data. The paper details the process of constructing the CHN-DF dataset and the Chinese data evaluation benchmark and validates the complexity and realism of the CHN-DF dataset through experiments. It is hoped that this evaluation benchmark will assist researchers in building more practical and effective facial video anti-forgery detection models, thereby advancing the technology in the field of anti-forgery detection. Additionally, this paper addresses the challenges posed by Chinese face video anti- forgery detection benchmark datasets and anti-forgery detection technology. It also proposes potential future research directions, offering valuable insights to advance the development of face video anti-forgery detection technology.

Key words: Deep Learning; Deepfakes; Fake Video; Multimodal Anti-Forgery Detection

* 基金项目:国家重点研发计划(2022YFF0902003); 宁波市自然科学基金(2022J182)

为了应对数字经济中数字智能面临的挑战,生成式人工智能(AIGC)[1]应运而生.通过基于用户输入的关键词或需求生成内容,AIGC 具有巨大潜力来支持不同应用.例如,依据当前维度的属性信息,AIGC 可以将数字内容从一个维度映射到另一个维度,实现对现实世界内容的智能增强和智能转译,从而极大地推动图像超分、语音转字幕以及文字转语音等自动化与执行效率.通过对当前内容的理解和属性控制,AIGC 可以修改视频内容,直接促进视频场景剪辑、虚拟试衣以及人声分离等视频内容理解技术的产业应用.在智能数字内容生成方面,AIGC 依托其从海量数据中学习抽象概念、通过概念的组合生成全新内容的能力,使得图像生成(AI 绘画)和视频生成的效果更加逼真.然而“眼见为实”的理念往往深植人心,在 AIGC 发展带动的视频内容生成变革情况下了解真相和信任这些信息变得越来越困难.AIGC 的发展无疑会造成人脸图片生成和人脸视频生成领域的信息真实性验证困难[2],对当今社会造成安全威胁甚至是挑战.例如,在 2022 年的俄乌冲突爆发阶段,网络上流传的乌克兰总统泽连斯基要求军队投降的视频片段以及俄罗斯总统普京宣布战争结束的深度伪造视频引发了双方国家甚至全球社会的恐慌.在 2023 年 4 月,美国共和党发布了 30 秒的深度伪造竞选广告,展示了一旦拜登赢得 2024 年竞选可能带来的灾难性场景.这类对政治人物的深度伪造视频通过形象抹黑和内容篡改,可能影响国家政治制度甚至引发国际战争危机.此外,社交身份的伪造导致各类诈骗现象不断增多.一些不法分子利用深度伪造技术塑造虚假的个人形象,在聊天室中通过面孔和声音模拟与“同龄”儿童进行数字对话,以获取未成年人的信任,从而对他们的安全构成威胁.

为了应对人脸视频深度伪造技术的滥用和潜在危害,工业界和学术界的大量研究人员提出了视频防伪检测技术[3-10].与此同时,人脸视频防伪检测数据集作为人脸视频防伪检测技术发展的基石,能够有效推动人脸视频防伪检测技术高质量发展.为了构建一个高效且可用的人脸视频防伪检测方法,需要大量多样化且高逼真的人脸视频防伪检测数据样本.因此,最近研究人员利用深度伪造方法创建了许多不同的人脸视频防伪检测数据集[11-20],旨在帮助研究人员训练和评估他们的视频防伪方法.然而,目前仍然缺乏用于训练多模态防伪方法的多模态深度伪造基准.现存的人脸视频防伪检测数据集大多数忽视了音频深度伪造和多模态深度伪造.虽然存在一些同时关注音频和视觉信息的多模态人脸视频防伪检测数据集,但在深度伪造的音频和视频方面通常存在数量和方法上的不平衡,且拍摄场景单一,这限制了视频防伪模型学习更一般性的多模态信息特征,进一步限制了视频防伪方法的发展.此外,现有的人脸视频防伪检测数据集主要集中在欧美人脸视频上,缺乏亚洲人脸视频数据样本,面向人脸视频防伪检测的大规模中文数据仍是空白.

为了弥补视频防伪数据集中多模态数据的缺乏和亚洲人脸视频数据样本不足,尤其是中文数据的空白,本文构建全球首个面向人脸视频防伪检测的大规模中文数据集—CHN-DF.CHN-DF 是最大的公开视频防伪数据集,样本量达到 426087.基于当前多种高逼真生成 AIGC 技术,CHN-DF 数据集覆盖了多样的取材场景并拥有庞大的视频数据样本量.数据源自 CN-CVS[21]与 CMLR[22],包含国内电视新闻和网络演讲节目中收集到的 2540 名说话人发言的视频片段,视频拍摄场景超过 2000 个,伪造视频则从音频与视觉信息两方面采用 Mockingbird[23]、FOMM[24]、FSGAN[25]、Motion-cos[26]、Simswap[27]、Wav2Lip[28]以及 coqui-TTS[29]总计 7 种主流深度伪造方法,以确保其内容足够复杂和多样化.为了搭建面向人脸视频防伪检测的评测基准,选用多模态视频防伪技术领域中主流的 11 种基线方法并对 CHN-DF 数据集进行综合实验,通过与人脸视频防伪检测领域已有数据集检测结果的对比,分析了现有防伪检测技术优劣与不足,验证 CHN-DF 数据集的多样性与实用性.

本文第 1 节介绍视频深度防伪数据集相关工作,第 2 节介绍数据集 CHN-DF,包括数据收集和生成.第 3 节介绍本文构建数据集的基准实验,通过实验结果验证了本文构建数据集的有效性,第 4 节介绍当下人脸视频防伪检测数据集与防伪检测技术面临挑战及发展方向,最后总结全文.

1 视频深度防伪数据集相关工作

AIGC 发展带来的视频内容生成技术变革,增加了检测人脸伪造视频的紧迫性,近些年来学术界和工业界的许多研究人员致力于创建人脸视频防伪检测数据集,开源了部分数据集以促进该领域的研究.本节将对人脸视频防伪检测数据集的现状进行梳理(见表 1).

表 1 视频深度防伪数据集汇总

数据集	类型	发布年份	真实视频数量	伪造视频数量	视频总数	说话人总数	伪造方法数量	真实数据来源
UADFV	视频	2018	49	49	98	49	1	YouTube
DeepfakeTIMIT	视频	2018	640	320	960	32	2	VidTIMIT
FF++	视频	2019	1000	4000	5000	未知	4	YouTube
Celeb-DF	视频	2019	590	5639	6229	59	1	YouTube
DeeperForensics	视频	2020	50000	10000	60000	100	1	演员拍摄
WildDeepfake	视频	2020	3805	3509	7314	未知	未知	网络收集
DFDC	视频+音频	2020	23654	104500	128154	960	8	演员拍摄
KoDF	视频+音频	2021	62166	175776	237942	403	6	演员拍摄
ForgeryNet	视频	2021	99630	121617	221247	5400+	8	VoxCeleb2 等
FakeAVCeleb	视频+音频	2022	500	19500	20000	500	4	VoxCeleb2
CHN-DF	视频+音频	2023	213187	212900	426087	2540	7	CN-CVS/CMLR

现有的人脸视频防伪检测数据集主要分为两类:第一类数据集借助视觉层面的单模态伪造方法,通过修改或交换人类的面部特征信息达到人脸伪造的效果;另一类数据集伪造方法结合视觉与听觉层面的伪造手段,对于一段真实视频,通过视觉或听觉特征信息的多模态修改实现视频信息的复杂伪造,此类伪造方法伪造角度与方式多样,更贴合人脸视频恶意伪造的现实情况,是视频深度防伪数据集的发展趋势.但要求伪造手段多样且过程复杂,因此此类数据集数据样本匮乏.

1.1 基于视觉的单模态人脸视频防伪检测数据集

- UADFV[11]:UADFV 为纽约州立大学研究人员在 2018 年发布的第一个用于人脸视频防伪检测的数据集,数据集共有 98 个视频,其中 49 个是从 YouTube 收集到的真实视频,伪造视频则是通过使用 FakeApp 应用程序[30]进行伪造生成出 49 个假视频.视频的平均长度为 11.14 秒,平均分辨率为 294×500 像素.作为早期人脸视频防伪检测数据集,UADFV 在数量和质量上都有限制,由单一的 FakeApp 产生的假视频中人脸扭曲变化及异常动作很明显,因此很容易检测到.
- DeepfakeTIMIT[12]:DeepfakeTIMIT 同样是在 2018 年引入的另一个针对深度伪造检测的人脸视频防伪检测数据集,该数据集的真实数据来源于 32 名说话人拍摄的 640 个视频,每个说话人视频集中包含 10 个高分辨率的 DeepFake-TIMIT-HQ 视频和 10 个低分辨率的 DeepFake-TIMIT-LQ 视频.假视频通过面部交换技术交换说话人间面部信息得到.然而,同样由于早期视频伪造方法的局限性,生成视频只有 4 秒长且合成的视频往往是模糊的.
- FF++[13]:FF+采用 4 种伪造手段 Deepfake[31],Face2face[32],Faceswap[33]和 NeuralTextures[34],是第一个假视频伪造方法既包含了基于深度学习的深度伪造方法,同时也涵盖了基于计算机图形学的伪造方法.数据集包含来自 YouTube 的 1000 个真实视频和 4000 个基于计算机图形学和两种基于深度学习的方法合成的伪造视频.此外,数据集划分成两个质量级别,即未压缩格式和 H264 压缩格式,可用于评估深度伪造检测方法在压缩视频和未压缩视频上的性能.然而,FF+的大小和多样性仍然不足,导致难以对由大量参数组成的高性能神经结构进行最优训练.
- Celeb-DF[14]:针对 UADFV、FF++和 DeepfakeTIMIT 等生成视频的质量不佳和篡改痕迹粗糙的问题,Celeb-DF 对视频伪造方法进行了改进,提供了更高质量的视频.数据集中的真实视频源自 YouTube 中的 59 位说话人的 590 个视频,并使用改进的 deepfake 技术生成了 5639 个虚假视频.然而,该数据集仍存在伪造方法单一的问题,不适用于现实世界中遇到的挑战.
- DeeperForensics[15]:数据集中的真实视频源自 100 名付费演员的录制,其中采用了 FF++中的视频作为面部交换伪造方法的 1000 个目标视频.通过将每个源身份与 10 个目标视频进行面部交换,合成了 1000 个假视频.此外,DeeperForensics 并没有采用其他的合成方法,而是利用 7 种扰动方法对真实视频和伪造视频进行数据增强以增加多样性.通过这种方式创建了 50000 个真实视频和 10000 个伪造视频.虽然数据量明显大于早期的人脸视频防伪检测数据集,并且更具多样性,但是 DeeperForensics 还没有像其他数据集一样对当前人脸伪造技术广泛的评测,因此 DeeperForensics 的学术效能尚未完全

建立.

- WildDeepfake[16]:面对早期人脸视频防伪检测数据集存在缺少内容多样性和视频源低质量的问题,WildDeepfake 从互联网上收集真实和深度伪造的样本,包含了视频中提取的面部动作序列,在人工去除没有对应真实人脸的视频后,真实视频数量为 3805,伪造视频数量为 3509.视觉效果更贴合真实生活场景,但数据量不足导致在训练高性能神经网络结构时存在局限.
- ForgeryNet[17]:目前为止是基于视觉的人脸视频防伪检测数据集中最大规模的数据集,提出了包括时序伪造定位、空间伪造定位等多项任务,ForgeryNet 采用 8 种深度伪造方法,生成 121617 个伪造视频.视频总量达到包含 221247,并且视频带有丰富的数据标注.

1.2 基于视觉与听觉的多模态人脸视频防伪检测数据集

- DFDC[18]:DFDC 是第一个在视频中包含伪造音频的数据集,起初作为 Facebook 发布的同名 DFDC 竞赛的数据集,包含 5250 个视频.之后经过数据补充真实视频达到 23654 个,伪造视频数据量达到 104500.为了保证数据集的多样性,真实视频源取自不同的环境设置,伪造视频则由八种不同的方法生成.听觉模态上仅进行音频交换,并没有使用音频伪造方法.标签仅包含真假两个类别,没有区别伪造视频中视觉伪造与听觉伪造.
- KoDF[19]:KoDF 是目前在基于视觉与听觉的多模态人脸视频防伪检测数据集领域中最大的公开数据集,包含采用 6 种伪造方法伪造的 175776 个假视频和 62166 个真实视频.视频中 403 说话人大多是韩国人,是为了平衡在现有的防伪数据集中亚洲人口数据不足的首次努力.然而 KoDF 在处理视觉与听觉信息时仅进行音频与人脸唇部动作的同步伪造,并没有使用声音克隆、声音转换等深度语音伪造方法.
- FakeAVCeleb[20]:首个同时包含伪造视频和伪造音频的人脸视频防伪检测数据集,是多模态人脸视频防伪检测常用的评测数据集,从 VoxCeleb2 数据集选择了 500 个真实视频,利用了 Faceswap,DeepFaceLab[35]和 FSGAN 伪造面部信息,利用 SV2TTS[36]伪造音频信息,使用 Wav2Lip 完成音频与人脸唇部动作的伪造,生成了 19500 个伪造视频.

2 CHN-DF 人脸视频防伪检测数据集

CHN-DF 人脸视频防伪检测数据集是首个面向人脸视频防伪检测的大规模中文数据集,该数据集包含视觉与听觉两个模态的信息.本节首先介绍 CHN-DF 数据集的真实视频获取和伪造视频生成,然后详细描述 CHN-DF 数据集的基本属性信息.

2.1 真实视频

为了保障 CHN-DF 数据集的场景多样性与内容复杂性,CHN-DF 真实视频源于目前最大的公开中文视听多模态数据集 CN-CVS 以及中文唇语数据集 CMLR.CN-CVS 总共有超过 2500 名说话人,数据总条数超过二十万,总时长超过 300 小时,CHN-DF 选取其中 Speech 部分的 2529 名说话人视频,选取的视频总量接近 20 万;CMLR 数据集包含了 2009 年 6 月至 2018 年 6 月的新闻联播视频,数据集包含由 11 位主持人所表述的共 102076 个视频,CHN-DF 数据集对 CMLR 数据集进行了筛选,达到保持说话人之间视频数据量平衡的目的,选取的视频总量接近 2 万.

基于此,CHN-DF 真实视频数据量达到 213187,超过目前公开的人脸视频防伪检测数据集的真实视频数量,说话人总数也达到 2540.此外,CMLR 使用基于 HOG 的人脸检测方法,再利用开源平台进行人脸识别和对齐;CN-CVS 使用 dlib 工具包对每个视频进行面部检测,删除没有人脸或多个人脸的视频.因此 CHN-DF 视频区域已固定在人脸部分.

由于 CHN-DF 数据集中真实视频基于说话人身份进行视频内容划分,数据集中训练集、验证集和测试集的说话人不存在重叠部分.因此,CHN-DF 数据集具有高度可扩展性.可以很容易地将新说话人的真实视频与伪造视频加入数据集,以增加真实和深度假视频的数量,并确保训练集、验证集和测试集相互独立.

2.2 伪造视频

CHN-DF 的伪造视频从音频与视觉信息两方面采用 Mockingbird、coqui-TTS、Wav2Lip、SimSwap、FOMM、Motion-cos 以及 FSGAN 总计 7 种深度伪造方法,覆盖主流的深度伪造方式.其中,Simswap 和 FSGAN 是基于面部交换的伪造方法;FOMM 和 Motion-cos 是基于面部重现的伪造方法;Mockingbird 和 coqui-TTS 是基于语音克隆的伪造方法;Wav2Lip 是基于唇语同步的伪造方法.图 1 显示了所选视觉伪造方法生成的示例,其中从上而下的每一行视频帧为依次使用 Wav2lip、SimSwap、FOMM、Motion-cos 和 FSGAN 创建的结果.不同方法伪造视频数量分布情况如图 2 所示,由于生成的伪造视频在人工检查过程中根据伪造效果进行了筛选,因此每种伪造方法的视频数量并不相等,但 CHN-DF 仍保持了伪造方法数量之间的相对平衡.此外,others 类别是指将源视频的音频替换为同一子集(训练集、验证集或测试集)下其他视频的音频后生成的伪造视频.



图 1 CHN-DF 伪造视频生成示例

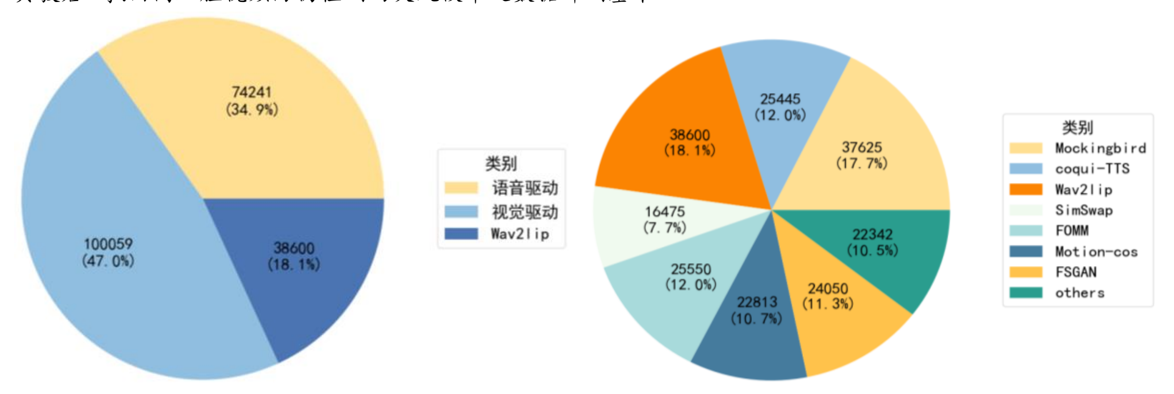


图 2 CHN-DF 中不同方法伪造视频数量分布

- Mockingbird:Mockingbird[23]用于中文实时语音克隆,通过不同讲话人音频信息合成虚假音频.在SV2TTS 的基础上,Mockingbird 引入中文训练数据集(aidatatang_200zh、magicdata、aishell3)用于训练语音合成系统,对训练数据集中的语音进行处理,提取讲话人的声音提取音色向量(Speaker Encode),然后根据讲话人声音和音色向量加上合成器(Synthesizer)和声码器(Vocoder)完成中文语音克隆.
- coqui-TTS:coqui-TTS[29]是一个低资源零样本文本转语音模型(Text-to-Speech,TTS),具有合成包括汉语在内的多种语言能力.提供了包括 Tacotron[37],Tacotron2[38],Glow-TTS[39]在内的多种文本语音规范模型,以及 MelGAN[40],Multiband-MelGAN[41],GAN-TTS[42]等声码器模型.这些模型的高效性和多功能性使得 Coqui-TTS 能够处理复杂的文本到语音转换任务,同时保持高质量的语音输出.
- Wav2Lip:Wav2Lip[28]是一个基于 GAN 的唇形动作迁移算法,Wav2Lip 不仅可以基于图片与目标语音匹配的唇形同步视频,还可以直接将动态的视频进行唇形转换,实现唇形动作与输入语音匹配的视频,即“对口型”.在原理上,Wav2Lip 利用预先训练的唇语同步检测器帮助模型根据音频学习嘴唇动作,实现生成的视频人物口型与输入语音同步.为了捕捉语音的时间上下文,该模型使用五个连续的人脸帧和对应的语音内容作为输入.
- SimSwap:SimSwap[27]模型采用身份注入模块(IIM),该模块可以在特征级别上将源图片中人脸的身份信息转移到目标视频的人脸上,此外使用弱特征匹配损失,该损失以隐式方式帮助模型保留面部属性.这些操作使得模型可以在实现通用且高保真度的面部交换.
- FOMM:FOMM[24]是作者通过自监督公式来解耦外观和运动信息的自监督模型,模型由运动估计模块和图像生成模块两个主要模块组成.根据目标视频中相似对象的运动,模型通过观察从同一视频中提取的帧对,将运动编码为特定于运动的关键点位移和局部仿射变换的组合,进而组合出学习运动的特征图来重建训练视频,应用时模型将源图像和目标视频的每一帧配对,并对源对象进行图像动画制作,从而实现生成关于源图像人脸的伪造视频.
- Motion-cos:Motion-cos[26]是一种用于部件分割的自监督深度学习方法,从人脸源图像中提取关键点信息,依据各个子部件的特征图对目标视频进行逐帧伪造,实现面部交换的区域化操作.Motion-cos 提供对人脸区域的五段、十段以及十五段分割预训练模型,CHN-DF 采用了十五段分割预训练模型对人脸区域进行细粒度的面部交换.
- FSGAN:FSGAN[25]是一种基于对抗生成网络的换脸模型,根据目标视频和源视频能够实现面部交换和面部重现.模型首先根据目标人脸的姿态和表情重新绘制源视频人脸并分割成两个面部区域,同时填补了重新绘制的脸部的缺失部分并将完整的脸部与目标进行混合,从而创造出最终的结果.在面部重现的过程中,模型通过 Delaunay 三角剖分选择与目标人脸最匹配的多个源视频人脸帧并使用重心坐标对再现结果进行加权平均,这个过程使得模型不需要为每个新源视频进行大量的调整.CHN-DF 采用了 FSGAN 中的面部交换技术.

2.3 数据集描述

2.3.1 数据集类别描述

使用上述深度伪造方法,CHN-DF 数据集基于视觉与听觉分为 4 个类别:真实视觉-真实听觉(V_RA_R)、真实视觉-伪造听觉 V_RA_F)、伪造视觉-真实听觉(V_FA_R)以及伪造视觉-伪造听觉(V_FA_F).

表 2 CHN-DF 数据集中视觉与听觉伪造组合类型与对方伪造方法

CHN-DF	真实听觉来源(A_R)	伪造听觉生成(A_F)
真实视觉来源(V_R)	数据源	Mockingbird,coqui-TTS
伪造视觉生成(V_F)	SimSwap,FOMM,Motion-cos,FSGAN	Wav2Lip,$V_F \times A_F$

(1)真实视觉-真实听觉(V_RA_R):V_RA_R 数据源自 CN-CVS 与 CMLR,从 CN-CVS 中选择 Speech 模块的 2529 名说话人视频,CN-CVS/Speech 具有大量的说话人和更加复杂多变的环境,贴合现实生活中对话场景和内容的复杂性;从 CMLR 数据集筛选近 2 万个 11 位主持人的主持视频.按照身份对出镜人编号,V_RA_R 数据总量达到 213187 个.

(2)真实视觉-伪造听觉(V_RA_F):V_RA_F 视觉上保持源视频的真实性,在听觉上进行音频伪造.如表 2 所示,在 CHN-DF 数据集中采用低资源零样本 TTS 模型 coqui-TTS 与基于迁移学习的中文实时语音克隆模型 Mockingbird 生成克隆的伪造音频.具体地,将源视频说话人的文本语句和其他说话人的音频作为模型输入,生成基于他人音频克隆的伪造音频.将伪造音频与源视频合并得到 V_RA_F 类别视频,这种类别的深度伪造可能的现实场景是一个人通过模仿另一个说话人的说话信息来进行身份欺诈.因此可以用来训练防御语音欺骗攻击.V_RA_F 数据总量达到 63070 个.

(3)伪造视觉-真实听觉(V_FA_R):V_FA_R 视觉上进行人脸伪造,视觉上保持源音频的真实性.如表 2 所示,人脸伪造通常采用面部交换和面部重现方法,在面部交换方法上采用 Simswap 和 FSGAN 模型,将源视频中的人脸与其他说话人的人脸进行面部交换.在面部重现方法上采用 FOMM 和 Motion-cos 模型,将源视频中的人脸帧与其他说话人的视频作为输入,实现其他说话人视频中的面部动作应用到源视频人脸上的效果.将伪造视频与源音频合并得到 V_FA_R 类别视频.在现实场景中存在攻击者通过修改他人的面部动作或交换人脸来塑造一个并不存在的视频画面,因此使用这种类别的深度伪造数据可以用来训练防御身份欺诈技术.V_FA_R 数据总量达到 88888 个.

(4)伪造视觉-伪造听觉(V_FA_F):V_FA_F 既包含人脸伪造又包含音频伪造,结合 V_RA_F 与 V_FA_R 伪造方法的同时又使用了 Wav2lip(见表 2).具体地,V_FA_F 包含三种伪造方式,第一种伪造方式为将时长相近的伪造音视频进行音视频合并;第二种伪造方式为将时长相近的伪造音视频合并之后采用 Wav2lip 进行唇形动作同步,即“对口型”;第三种伪造方式为对 V_RA_F 中视频数据采用 Wav2lip 改变唇形动作.V_FA_F 类别视频是对 V_RA_F 与 V_FA_R 类别视频的整合,贴合现实场景中视听觉同时伪造的复杂场景.V_FA_F 数据总量达到 60942 个.

值得一提的是,V_RA_F、V_FA_R 以及 V_FA_F 中伪造视频过程提到的其他视频与源视频均在同一子集(训练集、验证集或测试集)下,这保证了训练集、验证集和测试集相互独立.

2.3.2 数据集属性描述

CHN-DF 数据集包含 426087 个人脸视频,说话人总数达到 2540 人.其中真实视频 213187 个,伪造视频 212900 个,CHN-DF 正负样本平衡.负样本 V_RA_F、V_FA_R 以及 V_FA_F 的数量分别为 63070、88888 以及 60942 种类别伪造视频(即 V_RA_F、V_FA_R 以及 V_FA_F)的数量近似.

根据说话人身份,按照 7:1:2 的比例将 CHN-DF 视频划分为训练集(1778 位说话人的 350679 个视频)、验证集(254 位说话人的 22685 个视频)和测试集(508 位说话人的 52723 个视频),CHN-DF 视频时长分布如图 3 所示,持续时间在 0.36-355.58 秒,贴合现实情况下视频时长长短不一的特点,平均长度为 5.12 秒.视频时长集中在 0-20 秒,其中 98.75%的片段小于 20 秒,99.94%的片段小于 50 秒.

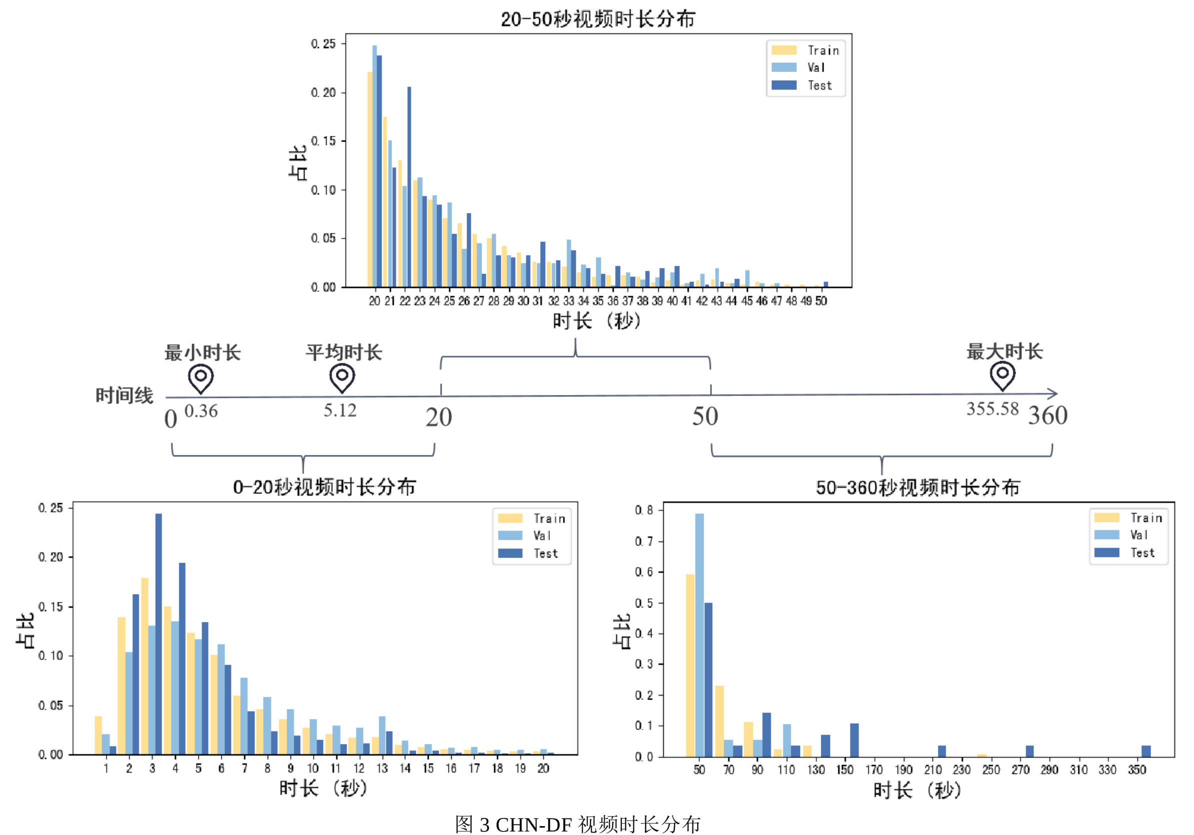


图 3 CHN-DF 视频时长分布

3 CHN-DF 基准评测

制作人脸视频防伪检测数据集的最终目标是推动研发出能够对各种深度伪造类型与方式表现良好的人脸视频防伪检测模型,人脸视频防伪检测模型性能好坏是通过测评模型在人脸视频防伪检测数据集的多种定量指标体现.在本节中将介绍 CHN-DF 基准评测的评估方法以及评价指标;基于代码的可复现性,采用 8 种多模态人脸视频防伪检测领域先进方法进行的全面基准性能评估,以此来展示 CHN-DF 数据集的复杂性和贴近真实场景水平,同时与最近发布的多模态 FakeAVCeleb 数据集进行比较.选择此数据集最重要的原因是 FakeAVCeleb 是目前已知的唯一包含详细音视频伪造标注的多模态人脸视频防伪检测数据集.此外,该数据集还采用了丰富的造假方法,在多模态人脸视频防伪检测领域是被广泛接受的优秀评测基准[43-51].

3.1 评估方法

在 CHN-DF 基准评测的评估方法选择中,按照数据集包含视觉与听觉两个模态信息的特点,选择基于单模态模型检测结果集成的防伪检测方法以及多模态人脸视频防伪检测模型进行基准评测.

3.1.1 集成方法

(1)Meso-4:Afchar[52]等人提出的四层卷积网络,是一种基于图像噪声中段信息的人脸伪造检测算法.这种方法有效解决了图像噪声减弱和高层语义特征难以区分伪造视频帧的问题.其浅层结构增强了对中等和大尺度特征的敏感度,提升了面部特征检测的能力.然而,这也带来了网络难以捕捉更深层次、更细微特征的局限.

(2)MesoInception-4:同样由 Afchar 等[52]提出.该模型架构的灵感来自于 InceptionNet[53],它通过用 InceptionNet 的模块替换第一层卷积层来改进 Meso-4,能够更有效地捕捉不同尺度上的特征.但也没能解决浅层网络结构在捕捉深层、细微特征方面的限制.

(3)Xception:由 Chollet[54]提出的一种完全基于深度可分离卷积层的卷积神经网络体系结构,对解耦通道相关性和空间相关性进行简化推导出深度可分离卷积,能够高效地提取图像和视频帧中的复杂特征.其复杂的网

络结构带来高效地特征提取能力的同时也可能导致训练和调整 Xception 模型变得更加困难.

3.1.2 多模态方法

(1)Multimodal-2:Multimodal-2[55]是一款开源的多模态模型,旨在预测电影类型,输入数据包括电影海报和类型.该模型由三部分组成:一部分是处理电影海报的卷积神经网络(CNN)块,负责视觉模式;另一部分是处理电影类型的长短期记忆(LSTM)块,负责文本模式;最后是一个前馈网络,负责分类,它综合了前两个模块的输出.在伪造视频检测中,模型利用 CNN 块分析视频帧的细微差异和 LSTM 块处理音频时序信息,有效捕捉伪造视频中的不一致性.

(2)CDCN:CDCN[56]基于中心差分卷积网络,用于解决人脸反欺骗的任务.该模型采用三层融合特征(低、中、高)来预测灰度面部深度.与传统的卷积神经网络相比,CDCN 通过其中心差分卷积网络能够有效地提取皮肤纹理、表情细节等细微的局部特征,有助于捕获伪造技术产生的微小瑕疵.

(3)MDS:音画同步是伪造视频难以伪造成功的,因为被伪造的视频帧往往会存在失去唇型或不自然的面部和嘴唇运动情况.MDS[57]比较伪造视频的视觉和听觉内容,通过量化模态之间的不协调性进行多模态伪造视频检测.

(4)VFD:VFD[58]关注人的生物特征(声音和面部)之间的匹配程度,利用了人类生物特征的内在相关性进行人脸防伪检测,学习面部和音频本质特征,拉近匹配的音视频,分离不匹配的音视频.

(5)AVoiD-DF:AVoiD-DF[59]是一种基于视听联合学习的人脸视频防伪检测方法,用于多模态人脸视频伪造检测.它由三个关键部分组成,包括时空编码器 TSE、多模态联合解码 MMD 和跨模态分类器 Cross-Modal Classifier,旨在通过深度伪造在时空层次上的视听不一致性进行伪造检测.

3.2 评估指标

为了评估人脸视频防伪检测模型在数据集上的性能优劣,本文采用准确度(Accuracy)、精确度(Precision)、召回率(Recall)和 F1 分数(F1-score)四项指标进行性能优劣的量化客观评估,不仅考虑到四项评估指标在分类领域使用的广泛性,而且在人脸视频防伪检测这一特定领域中,这些指标的组合利于全面评估模型性能、增强防伪问题场景的关注、处理类别不平衡、反映数据集质量.

(1)全面评估模型性能:Accuracy 衡量模型正确预测的总体比例,对于整体性能提供全面的视角,适用于平衡的数据集.F1-score 则将精确度和召回率结合,可以在正负样本之间取得平衡,在数据集存在类别不平情况下 F1-score 是一个综合的度量.

(2)增强防伪问题场景的关注:在人脸视频防伪检测这类安全防护场景中,更关心的是模型对真实正例的捕获程度,即模型的结果有多少是真正例,Precision 与 Recall 指标可以直观的反应人脸视频防伪检测模型结果在真正例结果上的优劣情况.

(3)处理类别不平衡:在人脸视频防伪检测模型中,人脸视频防伪检测数据集存在类别不平衡的情况,F1-score 适用于对模型性能评估时数据集类别不平衡的情况,可以更好地反映模型对正例的分类性能.

(4)反映数据集质量:凭借能够衡量模型正确分类比例的属性,Accuracy 对于正负样本平衡的数据集,可以作为对整体数据集质量的一个反映.在类别不平衡的情况下,F1-score 可以更敏感地反映模型对少数类别的处理能力,从而更好地评估数据集的质量.

3.3 基准实验与结果分析

3.3.1 数据集预处理

为了训练基准人脸视频防伪检测模型,分别按照视觉与听觉两个模态对数据集进行预处理.对于视觉模态,由于 CHN-DF 的数据源 CMLR 和 CN-CVS 视频区域已固定在人脸部分,因此无需进行视频的人脸检测与定位操作,从每个视频中提取视频帧并分别存储它们,然后将视频帧作为模型输入,提取用于分辨视频真伪的视觉特征.对于听觉模态,首先按照采样率为 16kHz 从视频中提取音频并以 WAV 格式存储.接着使用 10ms 窗口位移单位的 25ms 海宁(Hann)窗口计算梅尔倒谱系数(MFCC)特征.因此获得了每个音频帧包含 80 个 MFCC 特征的二维阵列(D=80),将所得到的 MFCC 特征存储为一个三通道图像,然后对 MFCC 特征图像的数量进行上采样来解

决每个视频只有一个 MFCC 特征图像的问题.将这些 MFCC 图像作为输入传递给模型,提取语音特征,以学习真伪语音之间的区别.

3.3.2 基准实验设置

为了 CHN-DF 基准评测的公平性,CHN-DF 中基准人脸视频防伪检测模型采用与评测 FakeAVCeleb 相同的模型参数.具体地,对每种方法进行了 50 次迭代的训练,使用了 EarlyStopping 机制,其中的 patience 设置为 10.采用了 Adam 优化器,学习率为 10^{-5},实验在一台搭载 Silver 4310 CPU 以及 Nvidia A40 GPU 的计算机上运行.其中在集成方法中,使用硬投票(Hard-Voting)和软投票(Soft-Voting)机制对音频和视频防伪模型进行预测结果投票集成[60].

3.3.3 多模态防伪方法对比实验

表 3 CHN-DF 数据集上多模态防伪方法对比实验

Methods	Year	CHN-DF				FakeAVCeleb			
		Acc.	Precision	Recall	F1-score	Acc.	Precision	Recall	F1-score
Meso-4(Soft-Voting)	2021	**0.5685**	0.4754	0.4729	**0.4741**	0.4593	**0.5373**	**0.5107**	0.3775
Meso-4(Hard-Voting)	2021	**0.4996**	0.5119	0.5096	**0.4793**	0.4593	**0.5373**	**0.5107**	0.3775
MesoInception-4(Soft-Voting)	2021	0.6455	0.7117	0.6541	0.6816	**0.7287**	**0.7445**	**0.7419**	**0.7286**
MesoInception-4(Hard-Voting)	2021	0.5811	0.5823	0.5337	0.5569	**0.7287**	**0.7445**	**0.7419**	**0.7286**
Xception(Soft-Voting)	2021	0.4360	**0.2686**	**0.5163**	**0.3533**	**0.4394**	0.2197	0.5000	0.3052
Xception(Hard-Voting)	2021	0.4360	**0.2686**	**0.5163**	**0.3533**	**0.4394**	0.2197	0.5000	0.3052
Multimodal-2	2021	0.5020	0.2510	0.5000	0.3342	**0.6740**	**0.6790**	**0.6735**	**0.6715**
CDCN	2021	0.5000	0.5000	0.5000	**0.4678**	**0.5150**	**0.5000**	**0.5004**	0.3855
MDS	2020	0.5784	**0.8571**	0.4521	0.5919	**0.6900**	0.7800	**0.6950**	**0.6650**
VFD	2022	0.6439	0.7113	0.6544	0.6816	**0.8152**	**0.8377**	**0.7542**	**0.7937**
AVoiD-DF	2023	0.6457	0.7244	0.6785	0.7006	**0.8371**	**0.8411**	**0.7702**	**0.8040**

本文选择 FakeAVCeleb 作为对比数据集,原因如 1.2 所述,在基于视觉与听觉的多模态人脸视频防伪检测数据集中 FakeAVCeleb 是目前已知唯一公开可获得的同时拥有深度伪造音频和深度伪造视频的数据集.因此为了进行面向人脸视频防伪检测的基准评测模型性能分析和通过实验验证 CHN-DF 数据集的复杂性和贴近真实场景水平, 将所选方法在 CHN-DF 与 FakeAVCeleb 上进行基准评测,性能结果如表 3 所示,两个数据集之间性能指标对比的最佳结果加粗表示.

3.3.3.1 CHN-DF 数据集的复杂性和贴近真实场景水平验证

11 种人脸视频防伪检测的基准评测模型(区分集成方法中硬投票和软投票机制)的共 44 项指标结果中,在 CHN-DF 中的指标结果共有 32 项低于在 FakeAVCeleb 中的指标结果.且基准人脸视频防伪检测模型在 CHN-DF 的四种指标结果集中在 0.6 以下.在侧重关注防伪问题场景的 Precision 与 Recall 指标,以及分别适用于正负样本平衡与失衡的 Accuracy 和 F1-score 上,人脸视频防伪检测基准评测模型在 CHN-DF 中性能相较于在 FakeAVCeleb 上表现不佳,面临的防伪任务更复杂、更具挑战性.由此也验证了 CHN-DF 数据集的复杂性和贴近真实场景水平, 更有利于推动性能更好的深度防伪检测方法研发.

3.3.3.2 面向人脸视频防伪检测的基准评测模型性能分析

AVoiD-DF 在 CHN-DF 和 FakeAVCeleb 中均取得了最佳性能结果,防伪效果最优.可能的原因是 AVoiD-DF 相较于其他基准人脸视频防伪检测模型,在基于视听联合学习模块引入的多模态联合解码 MMD 中,使用 MMD 模块进行模态融合.相较于其他多模态方法模态融合结构,AVoiD-DF 中输入的视觉和音频嵌入块是通过两个并行解码器通道馈送,每个通道都有一个双向交叉注意(BiCroAtt)模块,之后有自注意力块和前馈层.这使得两种模态之间具备更好的信息共享与联合学习能力.然而 AVoiD-DF 在 CHN-DF 的指标结果明显低于在 FakeAVCeleb 上的结果,可能的原因是 AVoiD-DF 作为基于视听联合学习的人脸视频防伪检测方法,在面对伪造视觉-伪造听觉(V_FA_F)情况时(如 Wav2Lip 将动态的视频进行唇形转换,实现唇形动作与输入语音匹配的视频)面部与音频的内在相关性会被破坏,同时 CHN-DF 相较于 FakeAVCeleb 在 V_FA_F 中采用更为复杂的伪造手段,因此 AVoiD-DF 在视听伪造信息更为复杂的 CHN-DF 中面对 V_FA_F 类别数据时指标结果较低;

MesoInception-4 在基于集成方法的人脸视频防伪检测基准评测模型中防伪效果最优,可能的原因是 MesoInception-4 针对伪造视频中伪造方法只能合成有限分辨率的人脸图像并且必须对其进行仿射变换以匹配源人脸的配置这一视频属性,使用变体 inception 模块关注仿射变换中扭曲面区域和周围环境的分辨率不一致

而产生的伪影.然而 MesoInception-4 在处理采用 FOMM 和 Motion-cos 等基于面部重现的伪造视频时,由于面部重现技术并不仅是将人脸区域进行仿射变换,面部重现更注重通过保留目标人物的身份来应用源人物的特征,而面部交换更注重在两个图像之间进行面部特征的交换.因此面部重现技术产生的伪影并不等同于面部交换过程中产生的伪造痕迹,MesoInception-4 在处理通过 FOMM 和 Motion-cos 生成的伪造视频存在局限性,导致指标结果较低;

Multimodal-2 与 Xception 在 CHN-DF 和 FakeAVCeleb 中指标结果较低,在 CHN-DF 中各项指标结果在 0.52 以下,造成这种结果的一个可能原因是 Multimodal-2 与 Xception 是计算机视觉领域通用分类模型,在各种分类任务中能够取得良好的结果,但可能是由于其预训练权重和特定任务之间的领域差异,而不一定适用于视频数据中的复杂特征和动态变化.另一方面,由于人脸视频防伪检测任务涉及到更丰富的信息,包括面部表情、姿势等因素,这可能导致了通用分类模型在该任务上的性能不佳.

此外,在面向人脸视频防伪检测的基准评测模型中多模态方法优于集成方法的性能结果,可能的原因是相较于集成方法中多个单模态分类器模型组成整体模型的思路,多模态方法在处理人脸视频伪造数据时考虑到视觉与听觉之间的相关性与一致性信息.相对于单模态(视觉或听觉)的伪造,伪造方法在篡改视觉与听觉之间相关性的特征时难度更大,使得伪造的效果更易于捕捉,所以视觉与听觉之间的相关性特征能够为人脸视频防伪检测模型提供更明显的检测特征,因此在处理具备多模态信息的人脸视频伪造数据中多模态方法防伪检测效果更优.

3.3.4 跨数据集防伪方法对比实验

为了评估 CHN-DF 数据集的质量和衡量基准人脸视频防伪检测模型的泛化性能,进行跨数据集防伪方法对比实验.实验使用基准模型在 FakeAVCeleb 进行训练并在 CHN-DF 上进行测试.通过在 FakeAVCeleb 上进行训练,模型能够学习人脸伪造视频的数据分布,在 CHN-DF 上进行测试能够提供模型在与训练集不同分布上数据中的性能表现.有助于验证模型在面对未知数据时的鲁棒性和泛化性,同时在 FakeAVCeleb 上的训练模型与在 CHN-DF 上的训练模型的测试结果对比也可评估 CHN-DF 数据集的质量.

表 4 跨数据集防伪方法对比实验

Methods	Year	Acc.	Precision	Recall	F1-score
Meso-4(Soft-Voting)	2021	0.4007	0.3844	0.4998	0.4345
Meso-4(Hard-Voting)	2021	0.4135	0.3321	0.4463	0.3808
MesoInception-4(Soft-Voting)	2021	0.4117	0.4100	0.4133	0.4116
MesoInception-4(Hard-Voting)	2021	0.4002	0.3911	0.4035	0.3972
Xception(Soft-Voting)	2021	0.3971	0.2134	0.4299	0.2852
Xception(Hard-Voting)	2021	0.3971	0.2134	0.4299	0.2852
Multimodal-2	2021	0.4145	0.3423	0.3997	0.3687
CDCN	2021	0.3784	0.3312	0.4521	0.3823
MDS	2020	0.5223	**0.6487**	0.4033	0.4973
VFD	2022	**0.6011**	0.5877	**0.5301**	**0.5574**
AVoiD-DF	2023	0.5997	0.6003	0.4983	0.5445

11 种人脸视频防伪检测模型在以 FakeAVCeleb 为训练集并以 CHN-DF 为测试集的跨数据集防伪任务中各项指标明显降低,表明模型在 CHN-DF 中面对了更复杂和更具挑战性的伪造数据,由此进一步验证了 CHN-DF 数据集的复杂性和贴近真实场景水平.

表 4 展示了跨数据集防伪方法对比实验结果,结合表 3 在 CHN-DF 数据集多模态防伪方法对比实验结果,可以发现由于数据集之间数据的来源不同,在跨数据集的防伪任务中 11 种人脸视频防伪检测模型性能指标有明显的下降.其中 MesoInception-4 指标结果下降最为显著,可能的原因是 MesoInception-4 在 FakeAVCeleb 中缺少基于面部重现的伪造视频的训练,导致通过捕捉伪影进行视频防伪检测的局限更加明显;VFD 在跨数据集的防伪任务中指标虽有下降但取得最优的防伪效果,可能的原因是 VFD 的微调(fine-tune)机制是基于预训练模型进行微调,因此可以快速适应新的任务或数据集;Multimodal-2、Xception 以及 MDS 在跨数据集的防伪任务中指标下降幅度较低,可能的原因是 Multimodal-2 与 Xception 作为通用分类模型虽然不一定适用于视频数据,但 Multimodal-2 与 Xception 良好的泛化性能使得模型在跨数据集任务中指标波动幅度降低.

4 面临挑战与发展方向

伪造与伪造检测是相互对立与辅助关系的复杂关系,为了应对快速发展的人脸视频伪造技术,人脸视频防伪检测技术也取得了长足的发展.然而,随着 AIGC 迅速发展,伪造技术已经可以生成高逼真图像与视频,给人脸视频伪造检测技术带来较大的冲击.当前阶段,人脸视频伪造检测技术发展已经落后伪造技术发展一大步,如何精准检测伪造人脸视频面临着巨大的挑战.因此,贴合真实场景的人脸视频防伪检测数据集,对于研发防伪效果更优的检测模型是十分必要且重要的;此外,现有人脸防伪数据集以欧美人为主,国际上缺少中文数据防伪数据集,因此构建面向人脸视频防伪检测的大规模中文数据评测基准,对于深度防伪技术的发展有重要的推动作用.

4.1 基准数据集构建局限性分析

本文构建的首个面向人脸视频防伪检测的大规模中文数据评测基准,在真实性、多样性、准确性、对抗性等方面仍存在诸多挑战,如何针对这些挑战,构建更优质的数据评测基准数据集,对于推动深度防伪检测技术高质量发展有着重要的意义.基准评测数据集构建当前存在的主要局限如下:

(1)深度伪造技术局限:AIGC 的发展使得图像生成(AI 绘画)和视频生成效果更逼真,然而现有的 AICG 伪造技术在生成长视频方面仍存在少量问题:i)讲话人说话期间存在面部短暂性闪烁现象;ii)存在伪造面部区域边缘模糊的情况;iii)面部纹理的过度平滑或缺乏细节;iv)头部姿势移动或动作不自然;v)缺乏面部遮挡物,如眼镜、光照效果等;vi)对身体姿态或皮肤颜色一致性变化敏感,易造成身份泄露;vii)伪造视频缺乏自然的情绪和语气停顿,会出现呼吸急促、语气僵硬的现象.这种因伪造技术带来的瑕疵也同样是伪造检测技术需要关注学习的特征,但过度关注这些特征会导致伪造检测模型过拟合,在真实应用场景鲁棒性的不足.同样的,这些伪造视频的瑕疵,一定程度干扰了评测基准的准确性与客观性,但当前阶段很难避免,现有的生成技术生成结果很难保证整体自然度、流畅性与连续一致性等贴近真实场景特性.为降低生成技术不足导致的评测基准不客观,本文在构建评测基准时通过人工筛选过于明显瑕疵数据,减少低质量对伪造检测技术定量评估成效的干扰.人脸视频防伪检测.

(2)语音数据缺乏多样性:现阶段人脸视频防伪检测领域评测基准中缺乏语音数据,在语音数据多样性方面难以保证,语音伪造技术缺乏包含多种情感表达的语音数据,使得评测基准无法充分覆盖对情感检测的测试.在多样化文化背景的语音数据收集上也面临巨大挑战,尤其是在中文数据方面,中文作为世界上使用人数最多的语言,涵盖多个方言和口音,而且不同地区和社会群体的语音表达方式各异.不同语音风格和口音数据的缺乏可能导致评测基准在应对特定口音或语音风格时的不足.因此,构建覆盖多样性、个性化的语音数据样本,也是本文未来工作的主要方向之一.

(3)标签缺乏准确性:有效的人脸视频防伪检测评测基准建立在能够贴近现实生活场景的数据集基础之上,贴近现实生活场景的数据集需要准确的标签,然而大规模的标注可能导致标签缺乏准确性,特别是在深度伪造的场景下,标注人员在标注视频数据时可能导致标签的主观性和不一致性,例如,对于 AIGC 创造的高质量伪造视频,人工标注会出现耗费大量时间但标签缺乏准确性的情况;在标注细粒度标签时,细粒度的标签需要标注人员对伪造技术有深入研究和专业知识,标注人员可能无法准确地识别所有细节.这些标签的主观性和不一致性情况会导致数据集在制作的过程中面临标签缺乏准确性的挑战.

(4)难以抵挡对抗性攻击:人脸视频防伪检测评测基准中缺乏对抗性攻击,现实场景中攻击者在制作伪造人脸视频的同时也会考虑如何加入对抗性攻击达到降低检测效果的目的,如刻意调整光线强度增加模型提取视觉特征的难度等等,导致训练出的防伪模型容易受到对抗性攻击的影响,这些复杂场景情况在人脸视频防伪检测评测基准的构建过程中难以有效考虑并覆盖到,导致在面对现实场景时伪造检测算法面临巨大的挑战.

4.2 人脸视频防伪检测技术面临挑战

人脸视频防伪检测评测基准与人脸视频防伪检测技术在攻防中互相促进、共同发展,构成人脸视频防伪检测领域的矛与盾.AIGC 的快速发展使得现有人脸视频防伪检测评测基准难以适应形势变化的同时,针对人脸视频防伪检测技术的研究同样面临诸多挑战.人脸视频防伪检测技术当前面临的主要挑战如下:

(1)大模型生成内容检测困难:人脸视频伪造技术发展之初,伪造视频中普遍存在视觉伪影或音频失真现象,

然而 AIGC 在视频内容生成的广泛应用,使得人脸视频伪造内容检测更加困难.以 ChatGPT 4.0 与 DALL-E 为代表的面向视频内容生成大语言模型的出现[61-63]使得人脸生成也随之迎来新一轮发展.凭借扩散模型通过训练神经网络来逆转添加高斯噪声带来的纯噪声,即从纯噪声中合成数据直到产生干净样本[64]的机制,使得人脸视频防伪检测技术难以捕捉视频伪造线索.给人脸视频伪造检测任务带来巨大挑战.

(2)难以应对复杂场景伪造内容:复杂场景的多样性增加了人脸视频防伪检测任务的复杂性.真实场景下人脸视频防伪检测工作易受环境因素干扰,例如,光照条件的改变可能使人脸的阴影和高光区域发生变化,使得人脸看起来更暗或更亮.摄像机角度的改变可能导致人脸的形状和特征发生畸变,使得人脸看起来扭曲或失真.此外,背景复杂性的变化也可能导致人脸的边缘模糊或与背景融合,使得人脸看起来不清晰或不成比例.以上这些因素都会对人脸视频的真实性和可信度产生影响,增加人脸视频防伪检测工作识别和检测的困难度.

(3)泛化性能差:现阶段针对单一人脸视频防伪检测数据集防伪检测技术防伪效果虽然较为理想,但在跨数据集防伪效果实验中泛化性能仍表现出明显的不足.同时在现实场景下由于面对人脸视频伪造方法未知,难以获得伪造方法的具体类型,因此在利用已有的人脸视频防伪检测预训练模型执行现实场景下的视频内容伪造检测任务时,检测结果可信度难以保证.

(4)防伪检测任务单一:目前人脸视频防伪检测任务侧重于对视频级伪造内容检测,检测任务粗糙.在现实场景下攻击者为了篡改视频信息内容,往往仅针对人脸视频中少数视频帧或少数音频段进行伪造,然而侧重于视频级伪造内容检测的防伪模型在面对大量视频帧或音频段中容易忽略伪造段特征信息,导致检测任务误判的概率大大增加.

4.3 发展方向

近年来,针对人脸视频防伪检测领域的研究已经取得显著的成果,但领域内依然存在诸多难点亟需解决.为了应对日益逼真的伪造人脸视频,本文聚焦人脸视频防伪检测技术与基准评测,为领域的发展提供新的视角与方向.在基准评测中可以从客观量化以及基准数据动态更新角度上思考;在防伪技术方面,可以从构建防伪自主进化机制以及注重防伪模型鲁棒性出发构思未来的发展方向.此外,促进人脸视频防伪检测发展的同时也应充分考虑数据隐私保护与社会影响.具体内容如下:

(1)评测基准客观量化:在 AIGC 技术发展带来的日益复杂、逼真视频内容伪造情境下,现有的评测基准却依赖于特定的模型性能评估指标结果,这会造成评测基准的角度局限性.因此在真实场景中,需要设计能够精准量化防伪模型多角度的防伪检测能力甚至是模型的自适应性能力,是未来评测基准构建的重要探索方向之一.

(2)基准数据动态更新:在设计评测基准时,需要充分考虑到复杂多样的人脸视频伪造类别的存在.因此定期更新基准数据集以纳入最新的伪造技术可以帮助评测基准贴近错综复杂的现实场景,考虑整合用户的反馈数据可以为基准数据集的动态更新提供新思路.此外,随着深度伪造技术的不断演进,建立动态的标签更新机制以应对新的深度伪造技术和生成模型也变得越来越重要.

(3)应对新型伪造检测:随着生成扩散模型、大模型等技术的快速发展,当前生成视频质量越来越贴近真实视频,以往针对合成伪造、生成模糊等特性的防伪方法,已经无法应对高逼真的生成视频.因此,针对新型伪造人脸技术,如何基于真实与伪造视频本身,以及样本在模型中局部特征相似性、模型推断路径等差异特性,设计相应防伪技术,是近几年亟需解决的难题.

(4)注重防伪模型鲁棒性:防伪模型是否具备强鲁棒性是能否在真实场景中应对复杂多变的视频内容保持稳定性和可靠性的关键.在防伪模型训练与测试过程中通过增加压缩率以及噪声干扰模拟真实场景分布变化,使防伪模型在构建过程中具备应对真实场景复杂多变视频的高鲁棒性;此外,在训练与测试过程中加入对抗性样本也是促进模型具备强鲁棒性的一种方式.然而,通过增加噪声、对抗样本等方式虽然一定程度能够增强模型的鲁棒性,但一定程度也能带来模型的识别性能损失,如何从真实样本本身特征出发,挖掘伪造样本与真实样本间差异,构建可应对任意伪造人脸视频的检测方法,且能够保证识别精准度,是未来的主要研究方向之一.

(5)自主进化防伪框架:伪造与防伪是相互对齐且相互促进技术,伪造技术的发展一般要领先于防伪技术一步,防伪技术在技术与性能上的落后,导致伪造人脸视频对人类社会带来较大的危害.当前的防伪模型与方法的设计主要依赖研究人员分析伪造技术的缺陷与不足,并针对性设计相应解决方案.如何针对变幻无常的伪造技

术,借助于对抗学习机制、强化学习模型等自主进化框架,设计能够推动防伪模型自主进化的框架,提高快速应对类型多变的伪造视频能力,是未来的重点的研究方向之一.

(6)考虑数据隐私保护:基准评测数据集构建与人脸视频防伪检测都需要充分考虑到对敏感信息的隐私保护.为此可以采用包括匿名化等技术手段在内的隐私保护手段,以确保在模型评测与应用过程中不会泄露用户的个人隐私信息.在推动技术进步的同时充分尊重用户的隐私权利,数据隐私保护是人脸视频防伪检测领域法制日趋完善过程中不可或缺的一部分.

(7)推动社会影响研究:目前人脸视频伪造领域缺乏完善的法律体系来实现对伪造视频的精准管控,例如区分具体的伪造视频是娱乐内容还是恶性传播,因此需要建立完善的法律体系对恶意制作或传播的互联网用户进行一定的惩戒[65].深入研究伪造视频对社会产生的潜在影响并进行社会伦理研究可以更全面地理解这一技术的社会影响并推动领域的可持续发展.这种关注不仅仅局限于技术层面,更需要注重在技术发展的同时充分考虑社会责任.

5 总 结

在AIGC时代人脸视频生成领域的信息存在真实性验证困难的环境下,本文提出了面向人脸视频防伪检测的大规模中文数据评测基准,发布了全球首个面向人脸视频防伪检测的大规模中文数据集——CHN-DF,填补人脸视频防伪检测数据集大规模中文数据的空白,本文详细介绍了构建 CHN-DF 数据集以及中文数据评测基准的流程,并针对主流防伪检测方法进行了对比实验,从基准评测模型性能、跨数据集泛化性能等方面分析了现有人脸视频防伪检测方法的优劣.此外,大量的实验也验证了 CHN-DF 数据集的复杂性和贴近真实场景水平,希望面向人脸视频防伪检测的大规模中文数据评测基准,能够帮助研究人员构建性能更为优异的人脸视频防伪检测模型,成为未来人脸视频防伪检测领域研究的基石.同时,本文还指出了中文人脸视频防伪检测数据集以及人脸视频防伪检测评测基准当前面临的挑战以及未来发展方向,希望为推动人脸视频防伪检测领域技术发展提供新的视角与方向.

References:

[1] Han Y, Li SY, Liu YX, Yan ZL, Dai YT, Philip S, Sun LC. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226. 2023.

[2] Zhao WX, Zhou K, Li JY, Tang TY, Wang XL, Hou YP, Min YQ, Zhang BC, Zhang JJ, Dong ZC, Du YF, Yang C, Chen YS, Chen Z, Jiang JH, Ren RY, Li YF, Tang XY, Liu ZK, Liu PY, Nie JY, Wen JR. A Survey of Large Language Models. arXiv preprint arXiv:2303.18223. 2023.

[3] Nguyen HH, Yamagishi J, Echizen I. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467. 2019.

[4] Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. In: Proc. of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: ICASSP, 2019. 8261-8265.

[5] Shao R, Wu TX, Nie LQ, Liu ZW. DeepFake-Adapter: Dual-Level Adapter for DeepFake Detection. arXiv preprint arXiv: 2306.00863. 2023.

[6] Haliassos A, Vougioukas K, Petridis S, Pantic M. Lips don't lie:A generalisable and robust approach to face forgery detection. In: Proc. of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: CVPR, 2021. 5037-5047.

[7] Chen HS, Rouhsedaghat M, Ghani H, Hu S, You S,.Kuo CC. DefakeHop: A light-weight high-performance deepfake detector. In: Proc. of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: ICME, 2021. 1-6.

[8] Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv: 2102.11126. 2021.

[9] Zhao H, Wei T, Zhou W, Zhang W, Chen D, Yu N. Multi-attentional deepfake detection. In: Proc. of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: CVPR, 2021. 2185-2194.

[10] Chen L, Zhang Y, Song Y, Liu L, Wang J. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. arXiv preprint arXiv: 2203.12208. 2022.

[11] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: Proc. of the 2019 IEEE Winter Applications of Computer Vision Workshops. Waikoloa: WACVW, 2019. 83-92.

[12] Korshunov P, Marcel S. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. arXiv preprint arXiv: 1812.08685. 2018.

[13] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. Faceforensics++: Learning to detect manipulated facial images. In: Proc. of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: ICCV, 2019. 1-11.

[14] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A new dataset for Deepfake forensics. arXiv preprint arXiv:1909.12962. 2019.

[15] Jiang L, Li R, Wu W, Qian C, Loy CC. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In: Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: CVPR, 2020. 83-92.

[16] Zi BJ, Chang MH, Chen JJ, Ma XJ, Jiang YG. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. arXiv preprint arXiv: 2101.01456. 2021.

[17] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang ML, Ferrer CC. The DeepFake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv: 2006.07397. 2020.

[18] Patrick K, Jaeseong Y, Gyuhyeon N, Sungwoo P, Gyeongsu C. KoDF: A Large-scale Korean DeepFake Detection Dataset. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal:ICCV, 2021. 10724-10733.

[19] He YA, Gan, B, Chen, SY, Zhou YC, Yin GJ, Song LCA, Sheng L, Shao J, Liu ZW. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Nashville: CVPR. 2021. 4358-4367.

[20] Khalid H, TariqS, Kim M, Simon S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv preprint arXiv: 2108.05080. 2022.

[21] Chen C, Wang D, Zheng TF. CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis. In: Proc. of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes: ICASSP. 2023. 1-5.

[22] Zhao Y, Xu R, ML Song. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. In: Proc. of the 1st ACM International Conference on Multimedia in Asia. New York: MMAsia '19. 2020. 1–6.

[23] Mockingbird. 2021 . https://github.com/babysor/MockingBird

[24] Siarohin A, Lathuili`ere S, Tulyakov S, Ricci E, Sebe N. First order motion model for image animation. In: Proc. of the Advances in Neural Information Processing Systems. Red Hook: NeurIPS. 2019. 7137–7147.

[25] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment. In: Proc. of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: ICCV. 2019. 7183-7192.

[26] Siarohin A, Roy S, Lathuilière S, Tulyakov S, Ricci E, Sebe N. Motion-supervised Co-Part Segmentation. In: Proc. of the 2020 25th International Conference on Pattern Recognition. Milan: ICPR. 2021. 9650-9657.

[27] Chen RW, Chen XH, Ni BB, Ge YH. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In Proc. of the 28th ACM International Conference on Multimedia. New York: MM '20. 2020. 2003–2011.

[28] Chung JS, Andrew Z. Out of time: Automated lip sync in the wild. In: Proc. of the Asian Conference on Computer Vision. 2017. 251-263.

[29] coqui TTS. 2023. https://github.com/coqui-ai/TTS

[30] FakeApp. 2019. https://www.deepfakescn.com

[31] Faceswap: Deepfakes software for all. 2020. https://github.com/deepfakes/faceswap

[32] Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. Face2face: Real-time face capture and reenactment of rgb videos. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. LasVegas: CVPR. 2016. 2387-2395.

[33] Faceswap. 2020. https://github.com/MarekKowalski/FaceSwap/

[34] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.

[35] Petrov I, Gao DH, Chervoniy N, Liu K, Marangonda S, Chris U, Dpfks M, Luis RP, Jiang J, Zhang S, Wu PY, Zhou B, Zhang WM. Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535. 2020.

[36] Jia Y, Zhang Y, Weiss R, Wang Q, Shen J, Ren F, Chen ZF, Nguyen P, Pang RM, Moreno I, Wu YH. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proc. of the 32nd International Conference on Neural Information Processing Systems. Red Hook: NIPS'18. 2018. 4485–4495.

[37] Wang YX, Skerry-Ryan R, Stanton D, Wu YH, Weiss R, Jaitly N, Yang ZH, Xiao Y, Chen ZF, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous R. Tacotron: Towards End-to-End Speech Synthesis. arXiv preprint arXiv: 1703.10135. 2017.

[38] Shen J, Pang R, Weiss R, Schuster M, Jaitly N, Yang ZH, Chen ZF, Zhang Y, Wang YX, Skerrv-Ryan R, Saurous R, Agiomvrgiannakis Y, Wu YH. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: ICASSP. 2018. 4779–4783.

[39] Kim J, Kim S, Kong J, Yoon S. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search. In Proc. of the 34th International Conference on Neural Information Processing Systems. Red Hook: NIPS'20. 2020. 8067–8077.

[40] Kumar K, Kumar R, Boissiere T, Gestin L, Teoh WZ, Sotelo J, Brebisson A, Bengio Y, Courville A. MelGAN: generative adversarial networks for conditional waveform synthesis. In the Proc. of the 33rd International Conference on Neural Information Processing Systems. Red Hook: NIPS. 2019. 14910–14921.

[41] Yang G, Yang S, Liu K, Fang P, Chen W, Xie L. Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech. In the Proc. of the 2021 IEEE Spoken Language Technology Workshop. Shenzhen: SLT. 2021. 492-498.

[42] Mikolaj B, Donahue J, Dieleman S, Clark A, Elsen E, Casagrande N, Luis CC, Karen S. High Fidelity Speech Synthesis with Adversarial Networks. arXiv preprint arXiv: 1909.11646. 2019.

[43] Zhang YB, Lin WG, and Xu JF. Joint Audio-Visual Attention with Contrastive Learning for More General Deepfake Detection. In the Proc. of the ACM Transactions on Multimedia Computing, Communications and Applications. TOMCCAP. 2023. Just Accepted

[44] Chao F, Chen ZY, Andrew O. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. In the Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: CVPR. 2023. 10491-10503.

[45] Liu XL, Yu Y, Li XL, Zhao Y. Magnifying multimodal forgery clues for Deepfake detection. Image Communication. 2023,118(C).

[46] Shahzad SA, Hashmi A, Khan S, Peng YT, Tsao Y, Wang HM. Lip Sync Matters: A Novel Multimodal Forgery Detector. In the Proc. of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai: APSIPA ASC. 2022. 1885-1892.

[47] Liu X, Yu Y, Li X, Zhao Y. MCL: Multimodal Contrastive Learning for Deepfake Detection. IEEE Transactions on Circuits and Systems for Video Technology. 2023. doi: 10.1109/TCSVT.2023.3312738.

[48] Cozzolino D, Pianese A, Nießner M, Verdoliva L. Audio-Visual Person-of-Interest DeepFake Detection. In the Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: CVPRW. 2023. 943-952.

[49] Yu Y, Liu X, Ni R, Yang S, Zhao Y, Kot AC. PVASS-MDD: Predictive Visual-audio Alignment Self-supervision for Multimodal Deepfake Detection. IEEE Transactions on Circuits and Systems for Video Technology. 2023. doi: 10.1109/TCSVT.2023.3309899.

[50] Ilyas H, Javed A, Malik KM. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing. 136(C). 2023. doi:10.1016/j.asoc.2023.110124

[51] Hashmi A, S. Shahzad A, Ahmad W, Lin CW, Tsao Y, Wang HM. Multimodal Forgery Detection Using Ensemble Learning. In the Proc. of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai: APSIPA ASC. 2022. 1524-1532.

[52] Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: A compact facial video forgery detection network. In Proc. of the 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong: WIFS. 2018. 1-7.

[53] Christian S, Liu W, Jia YQ, Pierre S, Scott R, Dragomir A, Dumitru E, Vincent V, Andrew R. Going deeper with convolutions. In Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: CVPR. 2015. 1-9.

[54] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: CVPR. 2017. 1800-1807.

[55] Verma D 2021. https://github.com/dh1105/Multi-modal-movie-genre-prediction

[56] Yu ZT, Zhao CX, Wang ZZ, Qin YX, Su Z, Li XB, Zhou F, Zhao GY. Searching central difference convolutional networks for face anti-spoofing. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: CVPR. 2020. 5295–5305.

[57] Chugh K, Gupta P, Dhall A, Subramanian R. Not made for each other-audio-visual dissonance-based deepfake detection and localization. in Proc. of the 28th ACM International Conference on Multimedia. 2020. 439–447.

[58] Cheng H, Guo YY, Wang TY, Li Q, Chang XJ, Nie LQ. Voice-Face Homogeneity Tells Deepfake. ACM Transactions on Multimedia Computing, Communications, and Applications. 2023. 20(3):1-22.

[59] Yang WY, Zhou XY, Chen ZK, Guo BF, Ba ZJ, Xia ZH, Cao XC, Ren K. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. IEEE Transactions on Information Forensics and Security, 2023, 18:2015-2029.

[60] Khalid H, Kim M, S, Tariq, Woo S. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proc. of the 1st workshop on synthetic multimedia-audio visual deepfake generation and detection. New York: ADGD '21. 2021. 7–15.

[61] Xi S, JX Ma, Zhou C, Yang ZX. Controllable 3D Face Generation with Conditional Style Code Diffusion. arXiv preprint arXiv: 2312.13941, 2024.

[62] Qing ZW, Zhang SW, Wang JY, Wang X, Wei YJ, Zhang YY, Gao CX, Sang N. Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation. arXiv preprint arXiv: 2312.04483, 2023.

[63] Zeng Y and Wei GQ, Zheng JN, Zou JX, Wei Y, Zhang YC, Li H. Make Pixels Dance: High-Dynamic Video Generation. arXiv preprint arXiv: 2311.10982, 2023.

[64] Ho J, Saharia C, Chan W, David J, Norouzi M, Salimans T. Cascaded Diffusion Models for High Fidelity Image Generation. arXiv preprint arXiv: 2106.15282, 2021.

[65] Li XR, Ji SL, Wu CM, Liu ZG, Deng SG, Cheng P, Yang M, Kong XW. Survey on Deepfakes and Detection Techniques. Ruan Jian Xue Bao/Journal of Software, 2021,32(2):496-518. (in Chinese with English abstract). http://www.jos.org.cn/jos/article/html/6140

附中文参考文献:

[65] 李旭嵘,纪守领,吴春明,刘振广,邓水光,程鹏,杨珉,孔祥维.深度伪造与检测技术综述.软件学报,2021,32(2):496-518. http://www.jos.org.cn/jos/article/html/6140